

DOCUMENT RESUME

ED 299 325

TM 012 360

AUTHOR Coffman, William E.
TITLE School and Teacher Effectiveness. Discussion.
PUB DATE 6 Apr 88
NOTE 8p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070) -- Viewpoints (120)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Research; Elementary Secondary Education; Higher Education; Measurement Techniques; *Research Methodology; Research Reports; *School Effectiveness; *Teacher Effectiveness; Test Reliability; Test Validity
IDENTIFIERS Criticism

ABSTRACT

Four papers presented at the National Council on Measurement in Education meeting are critiqued. G. R. Mandeville and H. Heidari (1988), in a study of unusually effective schools, asked if correlations among cohorts of schools differ depending on the method of analysis used. It is suggested that it would be better to ask whether schools identified as outstanding on the basis of data from one cohort would also be outstanding on the basis of data from other cohorts. R. F. Baglin (1988) proposed developing a short test to separate students into valid pass or fail groups. Such an objective, it is suggested, has undesirable consequences for many students. R. Hardy describes procedures used to develop measures of major field achievement at the college level, but does not indicate how to measure unique outcomes of distinct programs while providing a basis of comparison across programs. J. A. Burry and Shaw (1988) report the application of the Rasch Rating Scale Model to data from two parallel teacher rating scales. It is suggested that the analysis raises more questions than it answers. For these papers, it is appropriate to ask not only if the right questions were asked, but also if data analysis methods were appropriate. (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED 299325

DISCUSSION: SCHOOL AND TEACHER EFFECTIVENESS

William E. Coffman
University of Iowa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

WILLIAM E. COFFMAN

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

TM012360

National Council on Measurement in Education
Session D-2, April 6, 1988
New Orleans, LA

DISCUSSION: SCHOOL AND TEACHER EFFECTIVENESS

William E. Coffman
University of Iowa

The papers this afternoon deal with issues that have concerned evaluators for a long time. Mandeville & Heidari (1988) ask whether results will be more consistent from cohort to cohort within schools depending on the method of analysis that is used. Baglin (1988) asks whether or not it is possible to develop a short test that will separate students into valid pass or fail groups. Hardy (1988) describes procedures that are being used to develop measures of major field achievement at the college level that will reflect the unique nature of programs and at the same time provide some way of comparing results across programs and asks whether they are on the right track. Burry & Shaw (1988) ask whether or not an application of the Rasch Rating Scale Model to ratings of teachers will provide information more useful for diagnostic purposes than methods we have been using. These are obviously important questions. But a discussant might well ask two questions: (1) Have the investigators been asking the right questions given the situations they face? (2) Have they used the appropriate methods for analyzing their data?

The paper by Mandeville and Heidari (1988) begins by recognizing that if one is going to classify a school as unusually effective or worthy of special recognition, it is important to ask whether what one concludes from data collected and analyzed this year from a particular cohort of students can be generalized to other cohorts this year or to comparable cohorts in other years. They begin by noting that there seems to be a good bit of evidence that results are not the same from cohort to cohort and wondering whether this is a result of errors of measurement that might be reduced by better analytical methods. They then collect data from a number of different cohorts for a relatively large sample of schools in South Carolina and ask whether correlations among cohorts are different depending on the method of analysis. I believe that they have asked the wrong question and even for the question they did ask they have used the wrong method of analysis.

There isn't time this afternoon to give a detailed critique, but perhaps the following questions will stimulate additional thinking on their part. First, given that the focus was on the identification of unusually effective schools, why ask whether all schools in the sample tend to maintain the same relative rank from cohort to cohort. Wouldn't it be better to ask whether those few schools identified as outstanding on the basis of data from one cohort would also be identified as outstanding on the basis of data from other cohorts? Since most schools are likely to be mediocre and to differ primarily on the basis of random error, it may be that trying to demonstrate consistency on the basis of data from all schools in the sample might be giving the wrong answer. Second, why should one expect consistency for a single subject only? If teachers and administrators pay attention to test scores and base their teaching on test results, wouldn't you expect the teaching next year to focus on those areas where results were lower this year? It may be that the best teaching is that which does the most to foul up the predictions. In contrast, if achievement is the result of random accumulations, schools ought to maintain about the same overall achievement level depending on the input level, and change scores across schools ought to be random. In such cases, it might be that the better schools would be those with negative correlations between change scores this year and change scores next year. Third, couldn't one predict the results of applying the different analytic methods without bothering to carry out the analyses. If I read the theory correctly, predictions based on empirical Bayesian methods will surely produce predictions with smaller errors than those using classical methods. The question, however, is whether the reduction in error is worth the additional cost of analysis. Bayesian methods are very useful with very small samples, but the advantage reduces dramatically as the sample size increases.

Baglin (1988) proposes a criterion of a good school, i.e. one where at least 95% of the pupils are meeting some minimum academic mastery (MAM) at the end of each year of instruction. When I read the first draft of the paper I remembered what Dr. Robert Pace (1972) had said some years ago in a lecture in Iowa City. What he said was something like this: In evaluation the important question is not did we reach our objectives but rather what were the consequences. From what I know of the nature of individual differences and how children learn, as I indicated in my invited address this morning,

an objective like the one proposed by Baglin will inevitably have undesirable consequences for many pupils. As Walter Cook (1951) pointed out as long as 37 years ago, in schools where we are trying to educate all the children of all the people, retention of pupils can result only in an increase in individual differences within any grade in school and reduce the overall amount of learning. Periodically we rediscover this fact, and schools that have been concentrating attention on bringing slow learners up to some minimum become concerned about what is happening to the pupils who are more rapid learners. Thus, over the years we see a pendulum swing from concentration on minimum essentials to concentration on excellence and back again. I believe it would be well for Baglin to ask not only about MAM but also about what else is happening. It might actually be much better to see if one couldn't use tests to help teachers discover where in the learning sequence each pupil is and try to go on from there. When this is done, our data indicate that somewhere along the 12-years of schooling, most pupils do achieve some minimum level of achievement. Items that only a few pupils can answer correctly in the early grades are answered correctly by most pupils several years later. A single MAM cutoff based on some point on a longitudinal growth scale such as the ITBS GE scale which all pupils might be required to reach before leaving school would be much easier to defend than cutoff points for each grade.

Even if we accept the criterion of a promotable/retainable cutoff at each grade level, it is difficult to see how one might expect to develop a test as small as 25 items to do the job. The literature on attempts to develop cutoff points indicate that the results reported by Baglin could have been anticipated. The longer the test, other things being equal, the more reliable the classification. While the 25 item test chosen by a procedure that capitalized on chance did separate the criterion groups better than the longer test from which the items were selected, any attempt to apply the short test to a new sample is doomed to failure. This is a good example of the danger of using common sense in tackling a problem without studying the literature first. I know, for I made the same mistake myself before I learned about the importance of cross validation.

Hardy's (1988) paper does not attempt to provide answers to the question he and his colleagues have been facing, i.e. how to provide measures that

allow one to measure the unique outcomes of distinct programs while at the same time providing some basis of comparison across programs. Rather, this is an interim report that describes what actions are being taken and asks us to suggest modifications and/or additions to the procedure. There is no evidence in the draft of the paper I received to indicate the group has made any detailed study of previous attempts to develop measures that would be acceptable to different institutions or any search for records of previous experiences in developing test specifications or item pools. Perhaps this is due to the fact that ETS has accumulated its own experience; however, I'm not at all sure that the accumulated ETS experience is always communicated to new staff members. I think it might be a good idea for Hardy to ask directly about the ETS experience in developing the STEP tests (Educational Testing Service, 1959). I know that the effort to develop a classification scheme required months of effort and led to the publication of two books (French, 1957; Kearney, 1953), and that the workshops to develop the first drafts of the items were scheduled for a full week for each subject team. Is it possible that 3 days are not enough? Or that one ought to consider the possibility that it might be better to spend the original workshop period in orientation and training to be followed by a working period later to prepare drafts of test items?

It might be well also to consider whether or not there might be illustrative material from sources other than ETS. Sources that occur to me are the Manual of Examination Methods prepared in the 1930's by the University of Chicago Board of Examiners (1937); the report of the test development activity of the 8-year Study of the Progressive Education Association edited by Smith and Tyler (1942); the reports prepared by the staff of the Cooperative Study of Evaluation in General Education (Dressel & Mayhew, 1954); and early efforts of committees of examiners at ETS during the 1950's and 1960's leading to the first forms of the AREA Tests of the Graduate Record Examinations and the CEEB Tests of Developed Ability. It is my judgment that few if any current tests have achieved the quality of these pioneer efforts, probably because such test development is too expensive for large-scale application.

The other issue that I believe needs careful considerations is whether or not one ought to offer a fixed comparison block of items. The danger is that even though one cautions that this wouldn't be an equating set, there would be great pressure to so interpret it. Either major programs

are unique, in which case, comparisons are inappropriate except in terms of judgments regarding the relative value of the programs as offered, or institutions are competing by offering comparable programs, in which case, a common test is justified. One might be exercising one's responsibility as evaluation expert by insisting that the only defensible evaluation procedure is one that encourages within institution comparisons from year to year and leaves all across institution comparisons to judgments of panels of judges.

The final paper by Burry & Shaw (1988) reports the application of the Rasch Rating Scale Model to data from two parallel teacher rating scales, one consisting of self ratings of 211 teachers and the other of ratings of 204 teachers by their principals. The results of the scaling is reported in detail and the authors conclude thus: "The reliability coefficient of person separation is .98 on the observation form and .90 on self-report form. It appears that the Rasch Rating Scale Model is most appropriate for analyzing the IMP (Inventory of Myself as a Professional) on both instruments.."

If I am not mistaken, this conclusion means, simply: since these tests are reliable, they are valid. I recall that in the early years of applications of classical test theory, similar conclusions were common. You can find in the literature references to biserial correlations from item analysis results as "item validity coefficients," and it took some time for the idea to trickle down from the developers of theory that while a test needed to be reliable in order to be valid, a test could be highly reliable and not valid at the same time. I can still remember how surprised I was to find a validity coefficient for a test that was higher than the reliability even though the possibility was clearly implied in Gulliksen's equations.

One should be aware also of the fact that Rasch reliability is equivalent to an internal consistency reliability in classical test theory. I wonder, for example, what would be the correlation between the ratings of the 204 teachers by these particular principals and ratings of these same teachers by another set of raters. I don't think it would be anything like .98. One might also ask whether the same scale values would be generated if the three subscales of the IMP were scaled separately rather than as part of the total set of 36 items. The analysis reported here raises more questions than it answers.

References

- Baglan, R.F. (1988, April). An approach to the academic dimension in an effective school monitoring project. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.
- Burphy, J.A. (1988, April). Defining teacher effectiveness on a continuum. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.
- Cook, W.W. (1951). The function of measurement in the facilitation of learning. In E. F. Lindquist (Ed.), Educational measurement (pp. 3-46). Washington, DC: American Council on Education.
- Dressel, P.L. & Mayhew, L.B. (1954). General education: explorations in evaluation. Washington, DC: American Council on Education.
- Educational Testing Service. (1959). Cooperative Sequential Tests of Educational Progress: teacher's guide. Princeton, NJ: The author.
- French, W. (1957). Behavioral goals of general education in high school. New York: Russell Sage Foundation.
- Hardy, R. (1988, April). Item banking for higher education outcomes assessment. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.
- Kearney, N.C. (1953). Elementary school objectives. New York: Russell Sage Foundation.
- Mandeville, G.R. & Heidari, H. (1988, April). Measuring school effectiveness using hierarchical linear models. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.
- Pace, C.R. (1972). Thoughts on evaluation in higher education. Iowa City, IA: The American College Testing Programs.
- Smith, E.R., Tyler, R.W. & the Evaluation Staff. (1942). Appraising and recording student progress. New York: Harper.
- University of Chicago Board of Examiners. (1937). Manual of examination methods (2nd ed.). Chicago: University of Chicago Book Store.